An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano (rgiordan@mit.edu)[1]
January 2022

[1]With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

Dropping data: Motivation

More data & cheaper computation $\Rightarrow$
Statistical analyses are playing larger roles in decision making.

Decisions are important: We want **trustworthy** conclusions.
Data / models not always perfect: We want **robust** conclusions.

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Running example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit based on 16,560 data points.
We can reverse the studies qualitative conclusions by removing 15 observations ($< 0.1\%$ of the data).

How do we find sets of influential points? Difficult in general!

We provide a **automatic approximation** with finite-sample guarantees.

Studying the approximation reveals the causes of non-robustness.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Beta (SE)
Original result	-4.55 (5.88)

Original conclusion:
There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Original conclusion:
There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Original conclusion:
There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:
There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:
There is no evidence that microcredit is effective.

Potential conclusions after data dropping:
The effect of microcredit is positive (negative) & statistically significant.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:
There is no evidence that microcredit is effective.

Potential conclusions after data dropping:
The effect of microcredit is positive (negative) & statistically significant.

The culprit is signal to noise ratio.
By the end of the talk, we will see that the sensitivity is due to

- High variability of the outcome (hosehold profit) relative to
- A small signal driving the conclusion (statistical significance)

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always! But sometimes, surely yes.

Thinking without random noise can be helpful.

Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre. At harvest, you measure 200 bushels per acre.

- Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
  - Don't care about sensitivity to small subsets
- Scenario two: You want to recommend your farming methods to a friend across the valley.
  - Might care about sensitivity to small subsets

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Question 1:
How do we find influential datapoints?

Which estimators do we study?

Z-estimators. Suppose we have N data points $\vec{d} = d_1, \ldots, d_N$. Then:

$$\hat{\theta} := \vec{\theta} \text{ such that } \sum_{n=1}^{N} G(\vec{\theta}, d_n) = 0_P.$$

Examples: MLE, OLS, VB, &c (all minimizers of smooth empirical loss).

Function of interest. Qualitative decision based on $\phi(\hat{\theta}) \in \mathbb{R}$. E.g.:

- A particular component: $\phi(\theta) = \theta_d$
- The end of a confidence interval: $\phi(\theta) = \theta_d + \frac{1.96}{\sqrt{N}} \hat{\sigma}(\hat{\theta})$

Fix a proportion $0 < \alpha \ll 1$ of points to drop and find a set $\mathcal{S} \subset \{1, \ldots N\}$ with $|\mathcal{S}| \le \lfloor \alpha N \rfloor$ that extremizes $\phi(\hat{\theta})$ when dropped.

- **Problem:** There are many sets with $|\mathcal{S}| \le \lfloor \alpha N \rfloor$.
  - E.g., in Angelucci et al. [2015], $\binom{16{,}560}{15} \approx 1.5 \cdot 10^{51}$
- **Problem:** Evaluating $\phi(\hat{\theta}(\vec{d}_{-\mathcal{S}}))$ requires an estimation problem.
  - E.g., in Angelucci et al. [2015] computing the OLS estimator.
  - Other examples are even harder (VB, machine learning)

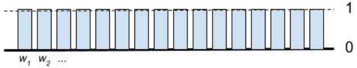
An approximation is needed!

Which estimators do we study?

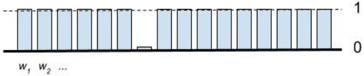
Suppose we have N data points $d_1, \ldots, d_N$. Then:

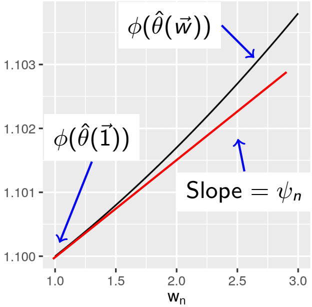
$$\hat{\theta} := \vec{\theta} \;\; \text{such that} \;\; \sum_{n=1}^{N} G(\vec{\theta}, d_n) = 0_P.$$

Original weights: $\vec{1} = (1, \ldots, 1)$



Leave points out by setting their elements of $\vec{w}$ to zero.





The slopes $\psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}$ are values of the **empirical influence function** [Hampel, 1986]. We call them "influence scores."

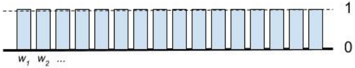
$\hat{\theta}(\vec{w})$ well-defined even for continuous values of the weights!

Which estimators do we study?

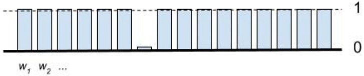
Suppose we have N data points $d_1, \ldots, d_N$. Then:

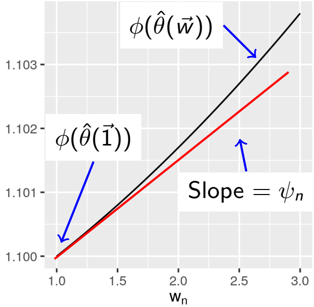
$$\hat{\theta}(\vec{w}) := \vec{\theta} \;\text{ such that }\; \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Original weights: $\vec{1} = (1, \ldots, 1)$



Leave points out by setting their elements of $\vec{w}$ to zero.





The slopes $\psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}$ are values of the **empirical influence function** [Hampel, 1986]. We call them "influence scores."

$\hat{\theta}(\vec{w})$ well-defined even for continuous values of the weights!

Taylor series approximation.

Problem: How large can you make $\phi(\hat{\theta}(\vec{w}))$ leaving out no more than $\lfloor \alpha N \rfloor$ points? **Combinatorially hard!**

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) \approx \phi^{\mathrm{lin}}(\vec{w}) := \phi(\hat{\theta}(\vec{1})) + \sum_{n=1}^{N} \psi_n (\vec{w}_n - 1)$$

Approximate solution: How large can you make $\phi^{\mathrm{lin}}(\vec{w})$ leaving out no more than $\lfloor \alpha N \rfloor$ points? **Easy!**

The most influential points for $\phi^{\mathrm{lin}}(\vec{w})$ have the most negative ψ_n.

Taylor series approximation.

Procedure:

Taylor series approximation.

Procedure:

1. Compute the “original” estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.

Taylor series approximation.

Procedure:

1. Compute the “original” estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \le \psi_{(2)} \le \ldots \le \psi_{(N)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \le \psi_{(2)} \le \ldots \le \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \le \psi_{(2)} \le \ldots \le \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \le \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \le \psi_{(2)} \le \ldots \le \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \le \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.
5. **Optional:** Compute $\hat{\theta}(\vec{w}^*)$, and verify that $\Delta \le \phi(\hat{\theta}(\vec{w}^*)) - \phi(\hat{\theta})$.

Computing the influence function.

How to compute $\psi_n := \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\Big|_{\vec{1}}$? Recall $\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P$.

Step zero: Implement software to compute $G(\theta, d_n)$ and $\phi(\theta)$. Find $\hat{\theta}$.

Step one: By the chain rule, $\psi_n = \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\Big|_{\vec{1}} = \frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta^T}\Big|_{\hat{\theta}} \frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\Big|_{\vec{1}}$.

Step two: By the implicit function theorem:

$$\frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\Bigg|_{\vec{1}} = \frac{1}{N} \left(\frac{1}{N} \sum_{n'=1}^{N} \frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_{n'}) \Big|_{\hat{\theta}} \right)^{-1} G(\hat{\theta}, d_n).$$

Step three: Use *automatic differentiation* on $\phi(\theta)$ and $G(\theta, d_n)$ from step zero to compute $\frac{\partial \phi(\theta)}{\partial \theta^T}$ and $\frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_n)$.

- The user does step zero. The rest is automatic.
- The primary computational expense is the Hessian inverse.
- Automatic differentiation is the chain rule applied to a program.
- Typically $\psi_n = O(N^{-1})$.

Question 2:
What makes an estimator non-robust?

Question 3:
When is our approximation accurate?

Conclusion:
Related work and future directions

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
"An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?"
https://arxiv.org/abs/2011.14999

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.